Biodiversity and Environment Data Mining

Somsack Inthasone, Nicolas Pasquier, Andrea G. B. Tettamanzi, Célia da Costa Pereira

ວາລະສານວິທະຍາສາດ

ຂອງ

ມະຫາວິທະຍາໄລແຫ່ງຊາດ

Scientific Journal of National University of Laos

ສະບັບ (Volume) ທີ 9

ພຶດສະພາ (May) 2015

ວາລະສານວິທະຍາສາດຂອງມະຫາວິທະຍາໄລແຫ່ງຊາດ

ຄະນະບັນນາທິການ

ວິຊາການຝ່າຍເຕັກນິກ

1. ຣສ. ດຣ. ໂພສີ ຈັນມີ່ງ

2. ອຈ. ຊະວະລິດ ແສງຊາດ

3. ນາງ ມະນີຍົມ ດອກດວງນາລີ

ຄຳນຳ

ທ່ານຜູ້ອ່ານທັງຫລາຍ, ວາລະສານວິທະຍາສາດຂອງມະຫາວິທະຍາໄລແຫ່ງຊາດສະບັບທີ 9 ນີ້ໄດ້ຖືກຈັດພິມຂຶ້ນໂດຍປະຕິບັດຕາມພາລະບົດບາດ ຂອງມະຫາວິທະຍາໄລແຫ່ງຊາດ ກ່ຽວກັບວຽກງານຄົ້ນຄວ້າວິທະຍາສາດ ແລະ ເພື່ອເປັນແຫຼ່ງຂໍ້ມູນກ່ຽວກັບຜົນການຄົ້ນຄວ້າຂອງຄູອາຈານ ແລະ ນັກສຶກສາໃນມະຫາວິທະຍາໄລແຫ່ງຊາດ.

ສະບັບນີ້, ບັນນາທິການ, ເຊິ່ງແມ່ນຄະນະກຳມະການປະສານງານການຄົ້ນຄວ້າວິທະຍາສາດຂອງມະຫາວິທະຍາໄລແຫ່ງຊາດ ໄດ້ຄັດເອົາຜົນການຄົ້ນຄວ້າວິທະຍາສາດ ຂອງອາຈານ-ພະນັກງານ ແລະ ນັກສຶກສາໃນມະຫາວິທະຍາໄລແຫ່ງຊາດ ຈຳນວນໜຶ່ງມາລົງ ແລະ ໃນນີ້ກໍ່ຍັງມີບົດທີ່ໄດ້ຮັບທຶນສະໜັບສະໜູນຈາກລັດຖະບານ ສປປ ລາວ, ສູນຄົ້ນຄວ້າອາຊີ ແລະ ຈາກແຫຼ່ງອື່ນໆ.

ນະໂອກາດນີ້, ພວກຂ້າພະເຈົ້າໃນນາມຄະນະບັນນາທິການວາລະສານວິທະຍາສາດ ຂອງມະຫາວິທະຍາ ໄລແຫ່ງຊາດຂໍສະແດງຄວາມຂອບໃຈມາຍັງນັກສຶກສາ, ນັກຄົ້ນຄວ້າ ແລະ ຄູອາຈານມະຫາວິທະຍາໄລແຫ່ງຊາດທີ່ໃຫ້ຄວາມຮ່ວມມືສົ່ງບົດຄົ້ນຄວ້າຂອງຕົນມາລົງວາລະສານສະບັບນີ້ ແລະ ພິເສດສະແດງຄຳຂອບໃຈນຳລັດຖະບານສປປລາວ, ສູນຄົ້ນຄວ້າອາຊີພ້ອມດ້ວຍກອງທຶນເກົາຫລີ ເພື່ອການສຶກສາຊັ້ນສູງທີ່ໄດ້ຮັບສະໜັບສະໜູນທຶນການຄົ້ນຄວ້າວິທະຍາສາດ ໃຫ້ແກ່ຄູອາຈານ, ນັກຄົ້ນຄວ້າຂອງມະຫາວິທະຍາໄລ ແລະ ທີມງານຈັດການພິມວາລະສານສະບັບທີ 9 ປະຈຳເດືອນຕົ້ນປີ 2015. ເຊິ່ງທ່ານສາມາດເຂົ້າເບິ່ງ ທີ່ເວັບໄຊ: http://www.nuol.edu.la. ພວກຂ້າພະເຈົ້າຫວັງຢ່າງຍິ່ງວ່າຈະໄດ້ຮັບຄວາມຮ່ວມມືຈາກບັນດາທ່ານໃນການປະກອບສ່ວນສົ່ງຜົນງານການຄົ້ນຄວ້າວິທະຍາສາດຂອງຕົນເອງ ລົງໃນວາລະສານສະບັບຕໍ່ໄປຂອງພວກເຮົາ.

ໃນວາລະສານສະບັບນີ້ບໍ່ອາດປາສະຈາກຂໍ້ບົກຜ່ອງ ທາງດ້ານຮູບການກໍຄືການນຳໃຊ້ພາສາວິຊາການ. ດັ່ງນັ້ນ, ຂ້າພະເຈົ້າຫວັງວ່າຈະໄດ້ຮັບຄຳຕຳນິສົ່ງຂ່າວຈາກບັນດາທ່ານຜູ້ອ່ານ ເພື່ອເຮັດໃຫ້ວາລະສານສະບັບຕໍ່ໄປມີຄວາມສົມບູນຂຶ້ນ.

(Forewords)

Dear Readers.

The ninth volume of the National University of Laos' (NUOL) Scientific Journal is established in accordance with the role of NUOL on research and dissemination which will be informative document for teachers and students of NUOL.

NUOL Research Coordination Committee, as the editors, have selected research papers conducted by teachers and staff, some of which were sponsored by the Lao-Government, and Asia Research Centre (ARC). Others are the independent research work which are outstanding.

On this occasion, we, on behalf of editors of this journal, would like to express our sincere thank to the researchers or NUOL staff for their good cooperation and great contributions, Lao Government and ARC for their financial support to the researchers, specially this journal.

We apologize for the short-coming, inappropriateness and linguistic errors which may occur in this publication. We are pleased to welcome all comments and feedback from all the readers of this publication for further improvement.

ສາລະບານ

Biodiversity and Environment Data Mining

Somsack Inthasone, Nicolas Pasquier,
Andrea G.B. Tettamanzi, Célia da Costa Pereira

Abstract

The preservation of biodiversity is a crucial challenge of our times. Computer Science has much to offer in support of the efforts made by scientists in a variety of disciplines to better understand and govern the phenomena related to biodiversity. We provide a survey of Information Technology based resources and methods that can be employed to this end. In particular, we discuss (i) sources of data and information available on the World Wide Web and the technologies that make them accessible and interoperable, (ii) the methods, collectively falling under the umbrella term of Data Mining, which allow us to discover new useful knowledge from the available facts and integrate expert knowledge in the process.

Keywords: Biodiversity, Data Mining, Information Technology, Ontologies.

1 Introduction

Biodiversity, a contraction of 'biological diversity', refers to the natural variety and diversity of living organisms [56]. Biodiversity is assessed by considering the diversity of ecosystems, species, populations and genes in their geographical locations and their evolution over time. The organization and distribution of ecosystems in bio-geographical scales also fall within the scope of this evaluation.

The *Earth Summit in Rio* held in 1992 under the aegis of the United Nations organization led to the definition of a global convention on biodiversity. All participating countries have officially agreed on the priority of protecting and restoring the diversity of life that is a vital resource for sustainable development. Indeed, some services provided by biodiversity are vital (oxygen and air, climate regulation, ecological balance, etc.) and the disappearance of key species or genetic diversity has unforeseen, and in most cases irreversible, consequences. Natural biodiversity is then an important factor of productivity and stability of ecosystems [37].

There is an agreement that the amount of information available increases every day. Most of that information is digital and then manageable with computer science techniques. While a few years ago, the problem was to find information, nowadays the problem is how to deal with an excess of information and to extract useful knowledge from it. This is true in many fields including biodiversity. Data mining provides theoretical frameworks, methodologies and techniques for analyzing large repositories of heterogeneous data [2]. In particular, there are currently numerous applications of data mining for extracting useful information and highlight new knowledge for decision making in various domains such as business, economy, scientific researches, and social and life sciences [1,7,18].

The aim of this article is to propose a survey of both data mining methodologies and existing information repositories which could help researchers in the biodiversity field in their intent to stop the decline of biodiversity and thus safeguard the diversity of life on earth.

Data mining helps to solve a number of important issues for biodiversity studies. Among these issues, we cite the research of phylotypes, the analysis and understanding of phenotypes and their evolution, and the analysis of genomics and proteomics data. Proposing a solution for solving these issues require analytical methods able to deal with large heterogeneous databases and geographical information systems storing the results of ecosystem studies and analyzes. Moreover, in order to improve the relevance and accuracy of extracted models, these methods should also allow to integrate both the available knowledge of the domain studied and the knowledge provided by the experts. The other aim of this article is therefore to propose a guideline for integrating available knowledge within the biodiversity data mining process.

The article is organized as follows. In section2 presents definitions, issues and challenges in the biodiversity fieldwork. Section3 presents literature on state-of-the art technologies and applications for biodiversity. Section4 briefly presents data mining approaches and their applications on biodiversity problems. Section5 concludes the article.

2 Biodiversity

There is a general agreement that biodiversity consists of three main types, or levels, of diversity as depicted in Figure1. The first type is *Genetic diversity*, the second one is *Species diversity*, and the last one is *Ecosystem diversity*, or Ecological diversity [20,48,50,56,57]. These are the three levels at which biological variety has been identified.

Figure 1: Types of diversity: Genetic (inner), Species (middle), and Ecosystem (outer)

Genetic diversity refers to the genetic variation and heritable traits within organisms. All species are related with other species through a genetic network, but the variety of genetic properties and features makes creatures different in their morphologic characteristics. Genetic diversity applies to all living organisms having inheritance of genes, including the amount of DNA per cell and chromosome structures. Genetic diversity is an important factor for the adaptation of populations to changing environments and the resistance to certain types of diseases. For species, a higher genetic variation implies less risk. It is also essential for species evolution.

Species diversity refers to the variety of living organisms within an ecosystem, a habitat or a region. It is evaluated by considering two factors: species richness and species evenness. The first corresponds to the number of different species present in a community per unit area, and the second to the relative abundance of each species in the geographical area. Both factors are evaluated according to the size of populations or biomass of each species in the area. Recent studies have shown relationships between diversity within species and diversity among species. Species diversity is the most visible part of biodiversity.

Ecosystem diversity refers to the variety of landscape of ecosystems in each region of the world. An ecosystem is a combination of communities - associations of species - of living organisms with the physical environment in which they live (e.g., air, water, mineral soil, topography, and climate). Ecosystems vary in size and in every geographic region there is a complex mosaic of interconnected ecosystems. Ecosystems are environments with a balanced state of natural elements (water, plants, animals, fungi, microbes, molecules, climate, etc.). Ecosystem diversity embraces the variety of habitats and environmental parameters that occur within a region. To preserve biodiversity, the conservation and protection of a representative array of interacting ecosystems, and their associated genetic and species diversities, is decisive.

Biodiversity applies wherever there is life, all around the world, from the earth's surface to marine ecosystems. Biologists most often define biodiversity as the "totality of genes, species, and ecosystems of a region". Ecologists consider biodiversity according to the three following interdependent primary characteristics: Ecosystems composition, i.e., the variety and richness of inhabiting species, ecosystems structure, i.e., the physical and three dimensional patterns of life forms, and ecosystems function, i.e., biogeochemical cycles and evolving environmental conditions. Even though many application tools have been developed, evaluating biodiversity still faces difficulties due to the complexity of precise evaluations of these parameters. Hence, the overall number of species that can be measured and officially identified all around the world is only 1.7 to 2 millions and 5 to 30 millions respectively [33,34].

2.1 Environmental Issues

In nature, biodiversity is the key to keep natural balance in changing environmental conditions. It functions as services, such as consumption service, that is to serve the natural resources to human (e.g., food, clothing, housing, medicines), industrial production service, that is to serve productivity of forest to be used either directly or indirectly (e.g., extracting chemicals from plants in the forest), and others (non-consumptive uses) including values of maintenance of ecosystems to be sustainable (e.g., soil maintenance, nitrogen to the soil, synthesis of plant power, humidity control) [15,19,52,57].

All life on the planet needs nutrients and oxygen, which are the main factors for survival. Especially, species depend on biodiversity resources produced by ecosystem services. The ecosystem services can regulate climate changes, dispose of wastes, recycle nutrients, filter and purify water, purify air, buffer against flooding, and maintain soil fertility [32,37]. Changes in environmental factors and ecosystems can thus endanger life forms as reported in several scientific studies.

The report of Global Biodiversity Outlook 3 [17] of the Convention on Biological Diversity (CBD) highlights Ban Ki-moon's speech, United Nations General Secretary, on the fact that "the consequences of this collective failure, if it is not quickly corrected, will be severe for us. Biodiversity underpins the functioning of the ecosystems on which we depend for food and fresh water, health and recreation, and protection from natural disasters. Its loss also affects us culturally and spiritually. This may be more difficult to quantify, but is nonetheless integral to our well-being". The loss of biodiversity becomes a serious issue for the twenty-first century. Its loss has direct and indirect negative effects on many factors connecting the elements of biodiversity as well as ecosystems [5].

In *Environmental factors*, the loss of biodiversity means that the natural balance between environmental conditions and the different types of diversities cannot be conserved, which will affect stability of ecosystems. This can lead to climate changes, such as the global warming reported in several scientific studies, and consequently to natural disasters (landslides, floods, typhoons, cyclones, hurricanes, tsunamis, etc.) [5,14,28,37].

Tourism factors are impacted by environment factors: If they are affected, by natural disasters or pollution for instance, tourism structure systems, such as aesthetic natural landscapes and historical places, can be affected, and even destroyed. For example, the effect of pollution on the structure and environment of the Venice city, in Italy, is known to endanger buildings [12,46,66]. This can impact tourism as natural landscapes and historical places are attraction sites for tourists, which consequently contribute to develop economical and social activities [5,28].

Human factors are linked to nutrients, oxygen, and other essential needs which are produced from biodiversity resources. If the number of biodiversity resources is decreased, the volume of vital products, such as food, water, plants and animals, will also decrease. This can lead to human's consumption and survivability concerns. For example, the augmentation of the production costs can lead to difficulties for human populations to have access to vital resources, such as medicine, food, etc. [4,5,6,8,28,32].

Society factors are affected by biodiversity loss as most parts of society infrastructures and livelihood depend on the basic system and structure of nature. One factor is nature productivity that depends on land structures for agriculture and irrigation, wood materials for building habitats, and natural and energy materials for other forms of consumption. For example, an important part of people living in rural societies depend mainly on productivity of agriculture and livestock for their livelihood, while people living in metropolitan areas need more biodiversity productivity as the demands for food, energy, materials and other resources are increased (transportation, construction, consumption products, etc.). Scarceness of resources can thus cause an augmentation in production costs leading to a reduction of the part of the population that has access to these resources [4,5,14,28].

Economics factors are impacted by direct benefits and added-values of natural resources (e.g., food, bio-fuels and renewable energies, animals and fibers, wood materials, bio-medical treatments). These resources contribute to the economic exchanges between countries around the world through internal and external commerce. Biodiversity loss, and scarceness of resources, can affect populations from an economical viewpoint. For instance, the important human population (more than 60 percent) that use bio-medication for main health cares [16]. It can also lead to higher production costs, implying more competition, financial crisis, and others economic related issues [5,6,27,38,44].

2.2 Topics and Challenges

Biodiversity loss is a major problem that bioscientists must take into account, considering and analyzing each parameter of loss. We describe below six main categories of causes and effects on biodiversity loss, as well as their impact on ecosystems and ecosystem services.

Habitat Loss and Fragmentation Habitat Loss are affected by many factors, for example, deforestation for agriculture, sawed timber, factories, etc. [27]. According to the *Global forest land-use change*

1990-2005 report by Food and Agriculture Organization (FAO) [31], the percentage of decrease in global forest areas is 1.7 percent in 30 percent (3.8 billion hectares) of overall forest areas around the world between year 1990 and 2005. This decrease is due to deforestation for agriculture, land uses, and other purposes. In [17], the authors used data and knowledge on percent of deforestation to address and warn about the problem of habitat loss. In addition, Habitat loss can be caused by natural disasters such as flooding, earthquake, landslide, and so on. However, habitats remaining from destruction are fragmented to small parts and resulting fragments are not enough wide for local organisms to live and migrate within, and among, other organisms [30,59].

Pollution Pollution of the air, land and water is caused mainly by human and natural factors, such as manufacture, transportation, construction, burned forest, electric power generation, and nuclear power generation [28]. This cause a risk of poisoning for all living organisms both on land and in the water on the planet. In addition, vehicle emissions, industrial emissions, and drainage of waste are factors that can increase carbon dioxide in the atmosphere and directly affect ecosystems. This can lead to climate changes, as well as global warming. According to *Global Health Observatory* report by the *World Health Organization* (WHO), the number of deaths by air pollution is about 4.6 million people in each year, and the worldwide percent of deaths by lung cancer about 9%, 5% of cardiopulmonary, and 1% of respiratory infection [6,53].

Climate Change The *CO_2Now Organization* provides a collection of global climate data from scientists around the world, showing the status of the global change [13]. These data show that the climate changes frequently occur and have different impacts on many aspects, especially regarding biodiversity, in each different zone of the world. For example, in the context of biodiversity in the Arctic zone the polar bears live on sea ice and other species living under sea ice are affected due to the elevation of temperatures in high latitudes [17,55]. On the other hand, the climate change is impacting life cycles of humans and other species on the earth, because of the more uncomfortable and unstable environment [53]. For example, the migration and adaption of human and other species to new locations due to frequent changes in environmental conditions [55]. However, the effect of climate change might give rise to abundance and distribution of individual species around ecosystems such as crops grow, breeding stock, the tides of the sea, etc. [4,44].

Invasive Alien Species This is the actual most important risk for biodiversity loss globally. Invasive Alien Species, whether present deliberately or coincidentally, can create intense issues in the biological ecosystem that can lead to the disappearance of numerous species and to difficulties to survive for other local species. The report of *Global Biodiversity Outlook 3* report of the *Convention on Biological Diversity* (CBD) depicts for each different category of species the different portions that are at distinct extinction risk levels. It is shown that among the 10,000 species listed, 2,000 species are in the risk or extinction zones. In [17], is also reported the sample data of alien species of 57 countries, where have been found more than 542 alien species. In addition, this issue will cause enormous investment expenses for farming, ranger services, fishery and other related human activities.

Human Populations The human population has a growing factor at an exponential rate and, according to the *United Nation report on World Population to 2300*, the size of the human population will increase from a growth rate of 2.3 to a growth rate of 36.4 billion. This will increase the consumption and may thus cause natural resources, as well as ecosystem services, to be insufficient. In order to preserve human life on the planet, the *Food and Agriculture Organization* (FAO) estimates and predicts data on the consumption of human population, and promotes the consumption of edible insects [60].

Overexploitation Overexploitation, in term of humans use of natural resources, means an over-consumption of ecosystem services by humans (e.g., in fisheries, hunting, and industries). This can lead to the destruction of the volume of natural resources, for example, the volume of fish stock decreased by fishery's overexploitation. According to the Food and Agriculture Organization (FAO) report, among the overall 600 marine fish stocks worldwide, 17 percent are overexploited, 52 percent are fully exploited, 20 percent are moderately exploited, 3 percent are underexploited, 7 percent are depleted, and 1 percent are recovering from depletion. Despite this small overexploitation percent (17%), this is an issue as natural resources and ecosystem services have limits to serve humans; the actual rate of resources consumption and the absence of natural resource protection causes risks for different species [19,60].

3 Resources and Technologies for Biodiversity

This section is devoted to the presentation of the major information providers and different resources available, and the technologies used to represent data and knowledge related to biodiversity and environment.

3.1 Resources

The most prominent information providers that propose data and knowledge repositories used for biodiversity and environmental studies are shown in Figure2. Each one provides contents related to different domains and categories depicted by the edges in the schema. These providers propose contents of different types (documents, databases, meta-data, spatio-temporal, etc.), in different categories (data, knowledge and documentations) and for different domains of application. Two main categories of resources are considered in this classification diagram. The *Data* category corresponds to resources depicting facts about species (animal and plants) and environmental conditions in specific areas. The *Frameworks* category corresponds to resources depicting both tacit and formal knowledge related to biodiversity and environment analytical application domains. Several providers, such as ESABII (East and Southeast Asia Biodiversity Information Initiative), IUCN (International Union for Conservation of Nature), and OECD (Organisation for Economic Co-operation and Development), give access to both data and frameworks resources.

In the *Biodiversity Policy* knowledge domain, information concern issues of principles, regulations and agreements on biodiversity. Amid these resources, we can cite BioNET, CBD, ESABII, IUCN and UNEP that provide information to serve and follow up among botanists, biologists and researchers in “globalization”. For example, the *Agreement of the United Nations Decade on Biodiversity 2011-2020* aims to support and implement the Strategic Plan for Biodiversity[1].

The *Environment* domain refers to results of researches and repositories on this area to scientists or people who want to know status of environment on the earth, especially the biologist who works on this issue. The major actors in this category are BHL, BioNET, CBD, ECNC, IUCN, KNEU and UNEP that are organizations which regularly publish reports and results of studies on domains related to biodiversity and environment, such as the *Protected Areas Presented as Natural Solutions to Global Environmental Challenges at RIO +20* published by the IUCN or the *Global Environment Facility* (GEF) published by the CDB organization.

The *Economics* domain corresponds to information from scientists about the status of economics development, based on effects and values of ecosystem services. The foremost information providers in this category are CBD, IUCN, OECD and TEEB. Among reports and studies published by these organizations, we can cite *Restoring World's Forests Proven to Boost Local Economies and Reduce Poverty* by IUCN and *Green Growth and Sustainable Development* by OECD for example.

In the *Health/Society* domain, repositories supply knowledge information referring to natural resources of ecosystem services and effects. These information have been produced, and their validity was demonstrated, by researchers from world-wide organizations such as BHL, CBD, IUCN, KNEU and UNEP. These organizations provide summaries and proposals such as for example, *Human Health and Biodiversity* is projected by CBD[2],*Towards the Blue Society* is projected by IUCN and *Action for Biodiversity: Towards a Society in Harmony with Nature* is published by CBD.

Information providers of data repositories and resources are categorized among the *Animals*, *Environment* and *Plants* categories depending on the research topics and domains of data they provide. These organizations are: FAOdata which web-based portal provides global data on biodiversity (i.e.,

1 http://www.cbd.int/2011-2020
2 http://www.cbd.int/en/health

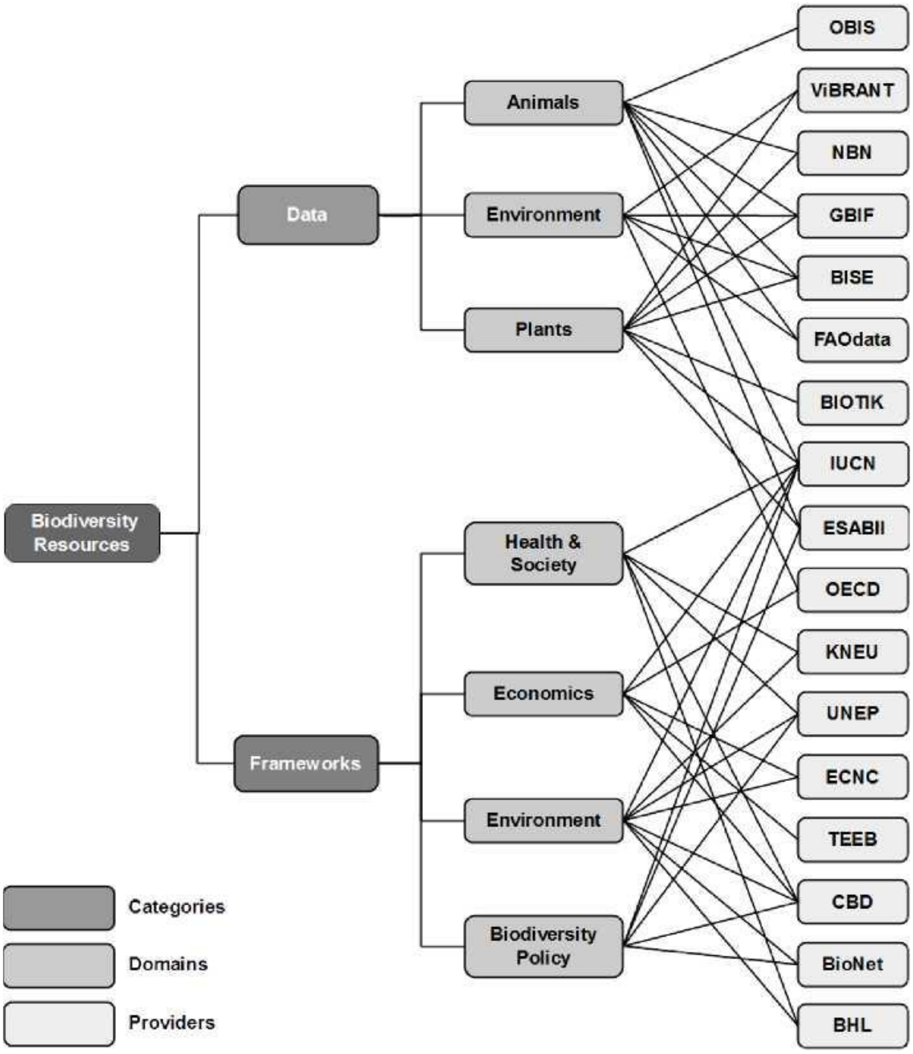

Figure 2: A classification diagram of biodiversity and environmental information

Data on Plants and Environment represented as datasets, statistics, spatial data, documents, images, etc.), BIOTIK that is a data repository on plants only in Southeast Asia and some other Asian countries, BISE that is a portal to serve data and datasets on biodiversity (Plants, Animals, and Environment) in European countries, GBIF that is a global data center (Data and Datasets) functioning as a hub of data collections (Plants, Animals, and Environment) from researchers around the world, NBN that is a portal to share biodiversity (Plants and Animals) data and datasets in United Kingdom, ViBRANT that is a web-based portal that aims to facilitate for research communities to merge and share biodiversity (Plants and Environment) data and datasets across European countries and some others, OBIS that is a web-based portal providing data on global marine species and visual spatial information on marine species from all the world's oceans (bio-geography).

3.2 Technologies

A very interesting technology that has been developed within the field of artificial intelligence as an outgrowth of early efforts aimed at representing knowledge consists of formal ontologies, which are a key for the semantic interoperability and integration of data and knowledge from different sources.

A definition of an ontology which makes justice of its complexity is the following, proposed in [22]:

> An ontology is a logical theory accounting for the intended meaning of a formal vocabulary, i.e., its ontological commitment to a particular conceptualization of the world. The intended models of a logical language using such a vocabulary are constrained by its ontological commitment. An ontology indirectly reflects this commitment (and the underlying conceptualization) by approximating these intended models.

In other words, an ontology may be regarded as "a kind of controlled vocabulary of well-defined terms with specified relationships between those terms, capable of interpretation by both humans and computers" [62]. From a practical point of view, an ontology defines a set of concepts and relations relevant to a domain of interest, along with axioms stating their properties. An ontology thus includes a taxonomy of concepts, a formally defined vocabulary (called a *terminology),* and other artifacts that help structure a knowledge base. In a sense, such artifacts may be considered as a generalization of the concept of metadata in database technology [47]. A knowledge base that uses the terms defined in an ontology becomes usable by and interoperable with any other system that has access to that ontology and is equipped by a logic *reasoner* for it [39].

Recently, an extensive standardization effort has been carried out by the World-Wide Web Consortium (W3C) in the framework of the Semantic Web movement. The Semantic Web is an extension of the WorldWide Web that enables people to share content beyond the boundaries of applications and websites [11]. The W3C has defined widely-accepted standards that make such an interoperability possible: the OWL 2 Web Ontology language defines the syntax that can be used to write ontologies; many reasoners are available today that are capable of using ontologies written in OWL 2 to make inferences on *facts* stored as RDF graphs [25]. A query language, SPARQL, is available for retrieving facts from RDF graphs in much the same way as data is retrieved from a database [51]. Data formatted using the RDF language and linked to ontologies are called *linked open data,* because their adoption of a standard format makes them usable to everybody and connected to all other data which refer to the same shared ontologies. Linked open data is the data layer of the Semantic Web.

Ontologies vary widely in scope and granularity. It is useful to distinguish four kinds of ontologies according to their level of generality [21]:

- *Top-level* or *upper* ontologies describe very general and fundamental concepts like space, time, matter, event, action, quality, etc., which are independent of a particular problem or domain.
- *Domain* and *task* ontologies describe, respectively, the vocabulary related to a generic domain (like biology) or a generic task or activity (like classifying or mapping), by specializing the terms introduced in the top-level ontology.
- *Application* ontologies describe concepts depending both on a particular domain and task, which are often specializations of both the related ontologies.

For example, a Plant Ontology (PO) [58], containing a conceptualization of plant structures (including plant cell, plant tissue, and sporophyte) and a controlled vocabulary for describing things like plant anatomy, plant morphology and plant development stage, may be described as a domain ontology.

In general, ontologies do not contain *facts*, i.e., data about the instances of the concepts they define and about their relationships or, when they do, these are limited to a few important facts that are useful to situate or organize the rest of the knowledge. Facts are usually stored in what we will call *fact repositories* which, in some cases, are implemented as or backed by a traditional RDBMS and may contain very large or huge amounts of data.

It is important to clarify that what may be considered an instance varies depending on the domain or task. When it comes to biodiversity, a plant species is generally treated as an instance, even though, strictly speaking, it should be regarded as a concept which groups together all specimens that share a set of phenotypic and genotypic characters. It is the specific application that dictates what should be treated as an instance or a concept.

In addition, ontologies have been extensively used in data integration approaches as they provide an explicit and computer-understandable conceptualization of a domain [10]. One of their major contributions to data integration and analysis is *mapping support* that is the use of an ontology of terms, formalizing a thesaurus, for the mapping process to facilitate its automation. The term mapping refers here to the semantical linking of data through the concepts represented in the ontology. Ontologies also provide capabilities of generalization and specialization of data according to the ontology concepts and their relationships. Each data can then be considered at the most appropriate level of abstraction, or aggregation, regarding the application objectives [29].

The Open Biological and Biomedical Ontologies (OBO Foundry) [3] is the portal of an ontology consortium that provides approaches and tools to help ontology development in different scientific domains of interest. In addition, the OBO Foundry is a large repository of candidate and validated ontologies in a number of scientific domains.

BioPortal [61] is a Web portal repository of open ontology resources that allows to search and explore ontology terms through an interactive visualization graphical user interface.

Ontobee [40,64] is a web tool to explore and browse ontological terms annotating linked data from different ontology repositories. Moreover, it supplies visualization and supports the SPARQL querying language, as well as management of data repositories of ontological terms, hierarchies and RDF format.

4 Data Mining for Biodiversity

4.1 Data Mining Concept

Data mining, also known as knowledge discovery from data (KDD), is a set of concepts, methods and tools for the rapid and efficient discovery of previously unknown information, represented as knowledge patterns and models, hidden inside massive information repositories [24]. The most prominent data mining approaches, gaining actually much importance in many application domains to support decisionmaking, are association rule and pattern mining, classification, clustering and regression. Since its emergence in the early 1990s, data mining made great strides and continues to flourish nowadays with the rapid evolution of automatic data acquisition systems, such as digital cameras, satellite remote sensing systems, bar code usage in retail, data streams in networks, text and image gathering tools, or Massively Parallel Signature Sequencing (MPSS) of genes for instance, and of storage systems, such as knowledge and data bases available on the World Wide Web, data warehouses and data marts, or publication repositories for instance. It is a multi-disciplinary field including solutions from database systems, statistics, knowledge-based systems, artificial intelligence, high-performance computing and data visualization.

A data mining process [35,36] is an iterative and interactive process that typically involves the three following general phases. During the pre-processing phase, data preparation techniques, i.e., data cleaning, integration, transformation and reduction methods, are applied to generate datasets containing relevant, consistent and reliable data, from the viewpoint of the application objectives, from heterogeneous data sources. The modeling phase consists in applying algorithmic methods for extracting knowledge patterns and models from the prepared datasets. During the post-processing phase, extracted knowledge patterns and models are presented to the end-user for interpretation and evaluation in order to discover novel information. The interactive and iterative nature of the process relies on the fact that changes and decisions made in the different steps of the three phases can result in changes in later steps and in the extracted patterns and models. Feedback loops between the phases are thus necessary to converge toward a satisfactory solution.

Even if considerable progress has been made in data mining since its early beginning, many challenges still remain. Hence, generic data mining systems can have limitations regarding application specific problems and a trend toward application dedicated systems can be observed nowadays [24]. These systems aim to handle complex heterogeneous data types, including multimedia, such as images, spatial and text data for instance, to generate complex knowledge patterns, to integrate domain specific knowledge represented both in knowledge bases and as users' methodologies and processes, to develop a unified theory of data mining would address problems in many fields, not only the biological and environmental one [24,65].

4.2 Techniques and Applications

Nowadays, many researchers and scientists attempt to solve biodiversity problems by using modern technologies, inventing approaches and techniques to measure and solve occurrence biodiversity issues. Because of the deluge of information found in data repositories of environmental sciences, coming from both institutions and amateurs, the use of data mining techniques for discovering new knowledge is spreading rapidly. There is a vast literature on data mining applications for biodiversity and environmental studies as follows.

In [9], the authors address the problem of the identification and classification of specimens through a knowledge-based discovery system. They extend the “classical” approach that consists of the three following phases: Grouping existing descriptions based on similarity measures by clustering, building and naming the classes identified by classification of groups, and reusing the formed concepts to identify the class of new observations. They combine inductive techniques and iterative neighbor search to take into account the structure (relationships between variables) and the content (missing, deviant and unknown values) of descriptions to improve the robustness of the classification. The resulting approach aims to help botanists and biologists to bring better evidences from knowledge based systems for the identification of new specimens through their observations. The proposed approach was validated with two knowledge bases built for coral classification and plants identification.

The problem of mountain biodiversity studies using data mining techniques is investigated in detail in [54] . It shows the importance of geophysical information systems for exploring and analyzing mountain biodiversity. The problems of the availability, quality and completeness of biodiversity data, that require consolidation before their use in analyses, are addressed thoroughly, demonstrating the importance of high-quality metadata, such as represented in ontologies and knowledge bases, for this complex task.

Several case studies, covering all biodiversity levels, from genes to species and ecosystems, using data mining techniques to analyze biodiversity patterns and processes along elevation gradients are reported. These show the relevance of data mining approaches for biodiversity conservation and protected area management, and the study of climate change effects on mountain biodiversity.

Biodiversity in forest ecosystems is the subject of the study in [41]. A data mining based approach is developed to predict biodiversity in forests by reasoning about their physical structure. The authors utilize a high-resolution scanning technology to capture different aspects of forests in three dimensional structure. These data are then related to the diversity of plants, invertebrates and birds in a range of forest types, to generate rich physical description datasets. These datasets are analyzed afterwards using five regression techniques from the popular Weka [23] data mining application. Results show that this approach can accurately predict six biodiversity measures of the species richness and the abundance of beetles, birds and spiders. This is a step toward the automation of the creation of a world forest inventory rich with environmental concerns.

In [26], the authors compare statistical and data mining methods for identifying relationships between a response and a set of predictors. They show that when little or no prior knowledge about the studied system is available, statistical models cannot accurately describe relationships between variations of the predictors and the response variable. Experiments were conducted using different datasets, including geographical, temporal, climate and species data, to compare results of six popular data mining tools with results of statistical techniques. The authors conclude that more use of data mining techniques should be made in environmental studies, whatever the degree of prior knowledge, and propose effective solutions to integrate data mining and statistical analyses into a thorough analysis.

The integration of geographical information from Geographical Information Systems (GIS) with species data, and its use in data mining studies is the object of the biodiversity informatics project of the W. P. Fraser Herbarium (SASK) [45]. The participants to this project develop an integrated bio-geography GIS model, using Google Maps API, based on data mining concepts to map and explore flora data. This research project shows that these data can be explored on a map and analyzed in several ways to reveal patterns showing relationships and trends that are not discernible in other representations of information.

The general problems of information integration and descriptive data quality are addressed in [43]. These problems are considered in the context of taxonomic classification of plant specimens into taxa, i.e., groups, according to the similarities between their observed features, or characters. This classification process relies mainly on the identification and description of variations between comparable structures of the different species. Since several terminologies and methodologies are in use for composing character descriptions, most often these descriptions are inconsistently composed, difficult to interpret and re-use, and data from diverse sources are not comparable. The authors propose a new conceptual model for unambiguously representing quantitative and qualitative description elements. It makes use of ontology technology to represent concepts and relationships in the descriptive terms. This model was implemented in a Java tool to help taxonomists to classify specimens and describe characters of new specimens through defined and controlled vocabularies.

In [49], the authors propose two approaches for representing plant classifications that are multiple and overlapping since in taxonomic classification of plants, some groups of specimens are referred to by a name used in different contexts over time. In both approaches, graph structures are used to represent classifications and relationships between them, but the two corresponding data models are different due to their constraints and aims regarding their capabilities to perform users' tasks. In the first approach, named Database Approach, the data model is dedicated to the storage of plant information. In the second approach, named Visualization Approach, the data model aims at the efficient retrieval and graphical exploration of plant information and their relationships. The results show that the two data models should coexist in a unique system for the automatic processing capabilities of the first one and the efficient exploration and comparison of classifications allowed by the second one.

Biodiversity applications showed that data mining can successfully discover new results and information to help environmental scientists to explain phenomena and get new insights in particular. However, these results can be improved by integrating data and knowledge from different related application domains into the biodiversity data mining process. For this integration, data, such as stored in application- level ontologies and databases, can be pre-processed using the structured representations of the domain knowledge stored in upper-level and domain-level ontologies. This approach can consolidate extracted information and help to solve complex problems of biodiversity and environmental studies that require to analyze data and knowledge from different domains together (e.g., environment, biology, geospatial topology).

5 Conclusions

Biodiversity refers to the variety and abundance of living organisms (plants, animals and other living beings) in a particular area or region. In an ecosystem, each species is part of the web of life and has a fundamental role in the circle of life. Hence, all species interact and depend upon one another for what each supplies, e.g., food, oxygen, shelter, and soil enrichment. Maintaining biodiversity of species in ecosystems is thus a necessity to preserve the web of life, and according to the biologist Edward O. Wilson, known as the “father of biodiversity”: “It is reckless to suppose that biodiversity can be diminished indefinitely without threatening humanity itself” [63].

Biodiversity loss is a major issue for all living species and preserving biological diversity in ecosystems requires to analyze and understand the parameters of this loss. This is a complex task for scientists as information from many domains (biology, geography, environment, pollution, etc.) must be considered and linked. This information can be categorized into two types: Knowledge, i.e., abstract concepts and relationships between them that can be general or specific to a peculiar domain and data, i.e., known and inventoried facts on concrete objects, which are described using knowledge concepts represented in ontologies. An important part of this information is available through web portals and repositories, but this information is most cases scattered, weakly documented and in formats that hinder their integration and analysis, and thus the discovery of new information. The definition of a methodology to integrate and structure data and knowledge into a unified information system that can serve as an integrated community resource is therefore a major concern for biodiversity and environment studies [42].

Data mining regroups theories, concepts and techniques for the analysis of large sets of weakly-structured heterogeneous data. The pre-processing, modeling and post-processing approaches proposed in this domain are thence adequate to both integrate and analyze biodiversity and environment data and knowledge. This is the core of the integration process to construct a biodiversity information system from which different datasets can be generated according to the specificities of the application or the analysis domain. This process allows to integrate and link information from different domains and from different types (e.g., text, images, spatial data) and to extract them together in data mining patterns and models without the requirement of peculiar treatments.

References

[1] S.J. Arfaee. Biodiversity prediction. Technical report, Department of Computing Science, University of Alberta, Edmonton, Canada, 2008.

[2] Thomas A.Runkler. *Data Analytics: Models and Algorithms for Intelligent Data Analysis.* Springer, 2012.

[3] The open biological and biomedical ontologies (obo).http://www.obofoundry.org, Accessed January 2013.

[4] J. Blanco and H. Kheradmand. *Climate change - Socioeconomic effects.* IntechOpen, 2011.

[5] B. Cardinale. Impacts of biodiversity loss. *Science,* 336(6081):552-553, 2012.

[6] S. Casalegno. *Global Warming Impacts - Case Studies on the Economy, Human Health, and on Urban and Natural Environments.* InTech, 2011.

[7] V. Chavan and P. Ingwersen. Towards a data publishing framework for primary biodiversity data: challenges and potentials for the biodiversity informatics community. *BMC Bioinformatics,* 10(Suppl 14):S2, 2009.

[8] COHAB. The importance of biodiversity to human health. *Biodiversity and Global Health*, 1, October 2010.

[9] N. Conruyt, D. Grosser, and R. Vignes-Lebbe. Knowledge discovery for biodiversity: From data mining to sign management. In R. Seppelt, A.A. Voinov, S. Lange, and D. Bankamp, editors, *Proceedings of the 6th International Congress on Environmental Modelling and Software Managing Resources of a Limited Planet,* Leipzig, Germany, July 2012. International Environmental Modelling and Software Society (iEMSs).

[10] Isabel F. Cruz and Huiyong Xiao. The role of ontologies in data integration. *Journal of Engineering Intelligent Systems,* 13:245-252, 2005.

[11] Michael C Daconta, Leo J Obrst, and Kevin T Smith. *The semantic web: a guide to the future of XML,*

web services, and knowledge management. John Wiley & Sons, 2003.

[12] M del Monte and Ottavio Vittori. Air pollution and stone decay: the case of venice. *Endeavour*, 9(3):117 - 122, 1985.

[13] Earth carbon dioxide (CO2).http://co2now.org, Accessed September 2013.

[14] EBI. Integrating biodiversity into environmental and social impact assessment processes. *Environmental and Social Impact Assessment*, Feburay 2013.

[15] N. Eldredge. *Life on Earth: An Encyclopedia of Biodiversity, Ecoloqy, and Evolution,* volume 1 of *Life on Earth*. ABC-CLIO, 2002.

[16] Kevin J. Gaston and John I. Spicer. Biodiversity: An introduction (second edition). *Oryx*, 38:465465, October 2004.

[17] Global biodiversity outlook 3.http://www.cbd.int/gbo3, Accessed January 2013.

[18] F. Gorunescu. *Data Mining: Concepts, Models and Techniques.* Intelligent Systems Reference Library. Springer, 2011.

[19] O. Grillo and G. Venora, editors. *Biological Diversity and Sustainable Resources Use.* InTech, 2011.

[20] Brian Groombridge and Martin Jenkins. *World Atlas of Biodiversity: Earth's living resources in the 21st century.* UNEP-WCMC, 2002.

[21] N. Guarino. Understanding, building, and using ontologies: A commentary to "Using explicit ontologies in KBS development", by van Heijst, Schreiber, and Wielinga. *International Journal of Human and Computer Studies,* (46):293-310, 1997.

[22] N. Guarino. Formal ontology and information systems. In N. Guarino, editor, *Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS),* pages 3-15, Trento, Italy, 1998. IOS Press.

[23] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations,* 11(1), 2009.

[24] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers Inc., San Francisco, USA, 3rd edition, 2011.

[25] Pascal Hitzler, Markus Krotzsch, and Sebastian Rudolph. *Foundations of Semantic Web Technologies.* Chapman & Hall/CRC, 2009.

[26] W. M. Hochachka, R. Caruana, D. Fink, A. Munson, M. Riedewald, D. Sorokina, and S. Kellings. Data-mining discovery of pattern and process in ecological systems. *The Journal of Wildlife Management,* 71(7):2427-2437, 2007.

[27] U. Karahalil and S. Keles. The effects of biodiversity concerns on economic profits of timber in forest management. In *Proceedings of the 7th Balkan Conference on Operational Research (BACOR),* Constanta, Romania, May 2005.

[28] Mohamed K. Khallaf. *The impact of air pollution on health, economy, environment and agricultural sources.* IntechOpen, 2011.

[29] Leonard Kwuida, Rokia Missaoui, Abdelilah Balamane, and Jean Vaillancourt. Generalized pattern extraction from concept lattices. *Annals of Mathematics and Artificial Intelligence,* pages 1-18, 2014.

[30] Gbolagade Akeem Lameed, editor. *Biodiversity Enrichment in a Diverse World.* InTech, 2012.

[31] E. J. Lindquist, R. D. Annunzio, A. Derrand, K. Macdicken, F. Achard, R. Beuchle, A. Brink, H. D. Eva, P. Mayaux, J. San-Miguel-Ayanz, and H. J. Stibig. *Global forest land-use change 1990-2005.* Food and Agriculture Organization of the United Nations and European Commission Joint Research Center, 2012.

[32] Millennium Ecosystem Assessment MA. Ecosystems and human well-being: biodiversity synthesis. *Washington, DC: World Resources Institute,* 2005.

[33] A.E. Magurran. *Measuring Biological Diversity.* Wiley, 2013.

[34] A.E. Magurran and B.J. McGill. *Biological Diversity: Frontiers in Measurement and Assessment.* Oxford biology. OUP Oxford, 2011.

[35] O. Marban, G. Mariscal, and J. Segovia. A data mining & knowledge discovery process model. In Julio Ponce and Adem Karahoca, editors, *Data Mining and Knowledge Discovery in Real Life Applications*. InTech, Vienna, Austria, 2009.

[36] G. Mariscal, O. Marban, and C. Fernandez. A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review,* 25(2):137-166, May 2010.

[37] G.F. Midgley. Biodiversity and ecosystem function. *Science,* 335(6065):174-175, 2012.

[38] P. Nijkamp, G. Vindigni, and P.A.L.D. Nunes. Economic valuation of biodiversity: A comparative study. *Ecological Economics,* 67(2):217-231, 2008.

[39] Leo Obrst. Ontologies for semantically interoperable systems. In *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM),* pages 366-369. ACM, 2003.

[40] Ontobee web portal.http://www.ontobee.org, Accessed February 2013.

[41] B. O'Sullivan, S. Keady, E. Keane, S. Irwin, and J. O'Halloran. Data mining for biodiversity prediction in forests. In *Proceedings of the 19th European Conference on Artificial Intelligence (ECAI),* pages 289-294. IOS Press, 2010.

[42] Cynthia S. Parr, Robert Guralnick, Nico Cellinese, and Roderic D. M. Page. Evolutionary informatics: Unifying knowledge about the diversity of life. *Trends in Ecology & Evolution,* 27(2):94-103, February 2012.

[43] T. Paterson, J.B. Kennedy, M.R. Pullan, A. Cannon, K. Armstrong, M.F. Watson, C. Raguenaud, S.M. McDonald, and G. Russell. A universal character model and ontology of defined terms for taxonomic description. In Erhard Rahm, editor, *Data Integration in the Life Sciences,* volume 2994 of *Lecture Notes in Computer Science*, pages 63-78. Springer Berlin Heidelberg, 2004.

[44] C. Perrings. Biodiversity, ecosystem services, and climate change - the economic problem. *Environmental Economics Series,* November 2010.

[45] C. Peters, D. Peters, and J.H. Cota-Sanchez. Data mining and mapping of herbarium specimens using geographic information systems: A look at the biodiversity informatics project of the W. P. Fraser Herbarium (SASK).http://www.herbarium.usask.ca/research/Data%20Mining,%20CBA%202009.pdf, 2009.

[46] R.K Pipe, J.A Coles, M.E Thomas, V.U Fossato, and A.L Pulsford. Evidence for environmentally derived immunomodulation in mussels from the Venice lagoon. *Aquatic Toxicology,* 32(1):59-73, 1995.

[47] Claudia Plant and Christian Bohm. *Database Technology for Life Sciences and Medicine*, volume 6. World Scientific, 2010.

[48] S. Popy. Definition des enjeux relatifs a la biodiversite en languedoc-roussillon. *Synthesis of the stakes related to biodiversity in Languedoc-Roussillon*, pages 3-20, 2009.

[49] C. Raguenaud, M. Graham, and J.B. Kennedy. Two approaches to representing multiple overlapping classifications: A comparison. In *Proceedings of the 13th International Conference on Scientific and Statistical Database Management (SSDBM),* pages 239-244, Fairfax, USA, July 2001. IEEE Computer Society.

[50] Osvaldo E Sala. (almost) all about biodiversity. *Science,* 299(5612):1521-1521, 2003.

[51] Andy Seaborne, Geetha Manjunath, Chris Bizer, John Breslin, Souripriya Das, Ian Davis, Steve Harris, Kingsley Idehen, Olivier Corby, Kjetil Kjernsmo, et al. Sparql/update: A language for updating rdf graphs. *W3C Member Submission,* 15, 2008.

[52] A. Shah. *Why Is Biodiversity Important? Who Cares?* Global Issues, April 2011.

[53] Raquel A Silva and et al. West. Global premature mortality due to anthropogenic outdoor air pollution and the contribution of past climate change. *Environmental Research Letters*, 8(3):034005, 2013.

[54] Eva M. Spehn and Christian Korner, editors. *Data Mining for Global Trends in Mountain Biodiversity.* CRC Press, 2009.

[55] Michelle D. Staudinger, N. B. Grimm, A. Staudt, S. L. Carter, F. S. Stuart, P. Kareiva, M. Ruckelshaus, and B. A. Stein. Impacts of climate change on biodiversity, ecosystems, and ecosystem services.http://assessment.globalchange.gov, July 2012.

[56] Ian R. Swingland. Biodiversity, definition of. *Encyclopedia of Biodiversity,* 1:377-391, 2001.

[57] J.A. Talent. *Earth and Life: Global Biodiversity, Extinction Intervals and Biogeographic Perturbations Through Time.* International Year of Planet Earth. Springer, 2012.

[58] The plant ontology (PO).http://plantontology.org, Accessed January 2013.

[59] Daniel H. Thornton, Lyn C. Branch, and Melvin E. Sunquist. The relative influence of habitat loss and fragmentation: Do tropical mammals meet the temperate paradigm? *Ecological Applications,* 21(6):2324-2333, August 2011.

[60] Arnold Van Huis, Joost Van Itterbeeck, Harmke Klunder, Esther Mertens, Afton Halloran, Giulia Muir, and Paul Vantomme. *Edible insects: Future prospects for food and feed security.* United Nations Food and Agriculture Organization (FAO), 2013.

[61] P.L. Whetzel, N.F. Noy, N.H. Shah, P.R. Alexander, C. Nyulas, T. Tudorache, and M.A. Musen. Bioportal: Enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Res.*, (39):W541-5, July 2011.

[62] P.L. Whetzel, N.F. Noy, N.H. Shah, P.R. Alexander, C. Nyulas, T. Tudorache, and M.A. Musen. What are ontologies?http://www.bioontology.org/learning-about-ontologies, Accessed March 2013.

[63] Edward. O. Wilson. *The Diversity of Life.* Questions of science. Belknap Press of Harvard University Press, 1992.

[64] Z. Xiang, C. Mungall, A. Ruttenberg, and Y. He. Ontobee: A linked data server and browser for ontology terms. In *International Conference on Biomedical Ontologies (ICBO),* pages 279-281, University at Buffalo, USA, July 2011.

[65] Q. Yang and X. Wu. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making,* 5(4):597-604, 2006.

[66] P. Zannetti, P. Melli, and E. Runca. Meteorological factors affecting SO2 pollution levels in Venice. *Atmospheric Environment,* 11(7):605-616, 1977.